METHODS AND APPARATUS FOR SLOW-STARTING A WEB CACHE SYSTEM

ABSTRACT OF THE DISCLOSURE

Methods and apparatus are described for intelligently assigning a portion of a cluster's traffic (e.g., buckets) to a cache system to minimize overloading of such cache system. In general terms, when a new cache system enters a cache cluster and/or starts up, the new cache system's full bucket allocation is not immediately assigned to the new cache system. Instead, only a portion of the full bucket allocation is initially assigned to the new cache system. Thus, the new cache system is less likely to be immediately overwhelmed as it enters a cache cluster. In one embodiment, the new cache system's bucket assignment is gradually increased until the cache system is handling it's full bucket allocation or it becomes overloaded. The cache system's load is also checked periodically (*e.g.*, every 30 seconds) to determine whether it has become overloaded. When the cache system becomes overloaded, buckets are immediately shed from the cache system. As a result, if the new cache becomes overloaded, it is unlikely to remain overloaded for a significant period of time. Thus, the new cache system is unlikely to cause a bottle neck for the cluster's network traffic. In sum, the new cache system's load is adjusted until it is handling an optimum number of buckets (*e.g.*, the cache is not underloaded or overloaded). In other embodiments, each cache system's load within the cache cluster continues to be monitored and adjusted so as to facilitate efficient use of each cache system.